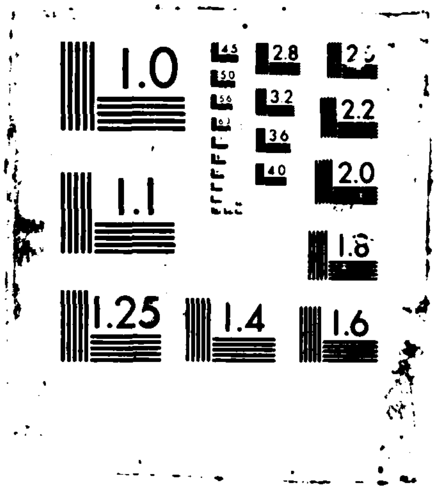
AD-A178 006 SOME ROBUST ESTIMATES OF COVARIANCE STRUCTURE BASED ON PARAMETRIC DENSITY..(U) RENSSELAUER POLYTECHNIC INST TROY NY SCHOOL OF MANAGEMENT A S PAULSON ET AL. 1987 1/1

UNCLASSIFIED WP-37-86-P27 ARO-18072.21-MA F/G 12/1 NL

END
4 87
DTIC


1.0
2.8
2.2
3.2
3.6
4.0
2.0
1.1
1.8
1.25
1.4
1.6

ARO 18072.21-MA

2

DTIC FILE COPY

AD-A178 806

Rensselaer Polytechnic Institute

School of Management

Working Paper No. 37-86-P27

87 4 1 2[illegible]4

SOME ROBUST ESTIMATES OF COVARIANCE STRUCTURE BASED ON PARAMETRIC DENSITY ESTIMATION

by

A. S. Paulson ,
T. A. Delehanty, and
N. J. Delaney

School of Management
Rensselaer Polytechnic Institute
Troy, NY 12180-3590
(518)266-6586

Note: This paper is not to be quoted without the permission of the author(s). For a list of Working Papers available from the School of Management, please contact Sheila Chao, School of Management.

A revised version will appear in Applied Statistics, Journal of the Royal Statistical Society, Series C.

SOME ROBUST ESTIMATES OF COVARIANCE STRUCTURE BASED ON PARAMETRIC DENSITY ESTIMATION

by

A. S. Paulson*

T. A. Delehanty

Rensselaer Polytechnic Institute

and

N. J. Delaney

Northeastern University

* Research sponsored in part by U.S. Army Research Office Contract DAAG29-81-K-0110.

SUMMARY

A new family of Fourier-based estimators of the parameters of the multivariate Gaussian distribution is presented. The estimators are equivalent to parametric density estimators. Three distinct estimators arise, each of which is robust and reduces to the maximum likelihood estimator as a special case. By varying the window width of a parametric density estimator, a set of diagnostics which are useful in problems of outlier detection and clustering are obtained. An example, using a trivariate data set, is given.

Covariance matrices.

1. INTRODUCTION

In fundamental papers Rosenblatt (1956), Parzen (1962), and Watson and Leadbetter (1963) considered the problem of estimating a density. While the approach used by these authors is non-parametric, we shall be concerned with density estimation in a parametric framework. In this framework the form of the density is assumed to be known apart from parameters which completely define the density. We shall exclusively consider the problem of estimating the parameters of the p-variate normal density

$$f(\underset{\sim}{x}) = |2\pi \underset{\sim}{D}|^{-\frac{1}{2}} \exp(-\tfrac{1}{2}(\underset{\sim}{x} - \underset{\sim}{\mu})^T \underset{\sim}{D}^{-1}(\underset{\sim}{x} - \underset{\sim}{\mu})) , \qquad (1.1)$$

where $\underset{\sim}{\mu}$ is the mean vector and $\underset{\sim}{D}$ is the covariance or dispersion matrix.

The characteristic function corresponding to (1.1) is given by

$$\phi(\underset{\sim}{u}) = \exp(i\underset{\sim}{\mu}^T \underset{\sim}{u} - \tfrac{1}{2} \underset{\sim}{u}^T \underset{\sim}{D} \underset{\sim}{u}) \qquad (1.2)$$

where $\underset{\sim}{u} = (u_1, u_2, \ldots, u_p)^T$ is a column vector of real numbers. If $\underset{\sim}{x}_1, \underset{\sim}{x}_2, \ldots, \underset{\sim}{x}_n$ is a random sample from (1.1), the sample characteristic function

$$\phi_n(\underset{\sim}{u}) = n^{-1} \sum_{j=1}^{n} \exp(i\underset{\sim}{u}^T \underset{\sim}{x}_j) \qquad (1.3)$$

is unbiased for (1.2) for every fixed $\underset{\sim}{u}$. Information concerning $\underset{\sim}{\mu}$ and $\underset{\sim}{D}$ can be extracted from $\phi_n(\underset{\sim}{u})$ by methods similar to those of Paulson and Nicklin (1983). In this case the estimators of $\underset{\sim}{\mu}$ and $\underset{\sim}{D}$

would be determined from score functions for $\underset{\sim}{\mu}$ and $\underset{\sim}{D}$ developed by differentiating

$$\int_{R_p} |\phi(\underset{\sim}{u}) - \phi_n(\underset{\sim}{u})|^2 \exp(-m^2 \underset{\sim}{u}^T \underset{\sim}{K} \underset{\sim}{u}) d\underset{\sim}{u} \qquad (1.4)$$

with respect to $\underset{\sim}{\mu}$ and $\underset{\sim}{D}$, setting $\underset{\sim}{K} = \underset{\sim}{D}$ and then setting the resulting expressions to zero. The expression (1.4) is therefore not a <u>bona</u> <u>fide</u> objective function.

An alternate approach is to proceed as follows. Let

$$\psi_n(\underset{\sim}{u}) = n^{-1} \sum_{j=1}^{n} \exp(i\underset{\sim}{u}^T \underset{\sim}{D}^{-\frac{1}{2}}(\underset{\sim}{x}_j - \underset{\sim}{\mu})) \quad . \qquad (1.5)$$

Whereas $\phi_n(\underset{\sim}{u})$ is a statistic, $\psi_n(\underset{\sim}{u})$ is not, but

$$E(\psi_n(\underset{\sim}{u})) = \psi(\underset{\sim}{u}) = \exp(-\tfrac{1}{2}\underset{\sim}{u}^T \underset{\sim}{u}) \quad . \qquad (1.6)$$

The parameters $\underset{\sim}{\mu}$ and $\underset{\sim}{D}$ of (1.5) may be estimated by making $\psi_n(\underset{\sim}{u})$ and $\psi(\underset{\sim}{u})$ match up in some sense. There are many ways in which this may be done, but we shall present only that which has been found to be most theoretically as well as computationally convenient and useful. Computational convenience is not a matter of small importance, it is in fact the overriding consideration when the dimension p becomes large; an estimator is effectively worthless if it cannot be calculated.

We shall ultimately obtain three pairs of estimators of $(\underset{\sim}{\mu}, \underset{\sim}{D})$, all arising from consideration of matching up $\psi_n(\underset{\sim}{u})$ with $\psi(\underset{\sim}{u})$. All pairs of estimators will be useful in the identification of potential outliers, in exploratory data analysis, and in clustering problems. Furthermore, all estimators will be qualitatively robust in the sense that their multidimensional influence functions will be bounded and re-descendant; all will reduce to familiar estimators as a special case.

2. DERIVATION OF THE ESTIMATORS BASED ON $\psi_n(\underset{\sim}{u})$

Define the function

$$Q_{m,n}(\underset{\sim}{\mu},\ \underset{\sim}{D}) = \int_{R_p} |\psi(\underset{\sim}{u}) - \psi_n(\underset{\sim}{u})|^2 \exp(-m^2\underset{\sim}{u}^T\underset{\sim}{u})d\underset{\sim}{u} \quad . \qquad (2.1)$$

The function $Q_{m,n}(\underset{\sim}{\mu},\ \underset{\sim}{D})$ is a <u>bona fide</u> objective function for $\underset{\sim}{\mu}$ and $\underset{\sim}{D}$ when $0 < m < \infty$. The integrand of equation (2.1) can be rewritten in terms of the residuals in $\underset{\sim}{u}$,

$$R(\underset{\sim}{u}) = \exp(-\tfrac{1}{2}(1 + m^2)\underset{\sim}{u}^T\underset{\sim}{u}) - n^{-1}\sum_{j=1}^{n}\exp(i\underset{\sim}{u}^T\underset{\sim}{D}^{-\frac{1}{2}}(\underset{\sim}{x}_j - \underset{\sim}{\mu}) - \tfrac{1}{2}m^2\underset{\sim}{u}^T\underset{\sim}{u}) \quad (2.2)$$

as

$$Q_{m,n}(\underset{\sim}{\mu},\ \underset{\sim}{D}) = \int_{R_p} R(\underset{\sim}{u})\ R^*(\underset{\sim}{u})d\underset{\sim}{u} = \int_{R_p} |R(\underset{\sim}{u})|^2\ d\underset{\sim}{u} \ ,$$

where $R^*(\underset{\sim}{u})$ denotes the complex conjugate of $R(\underset{\sim}{u})$.

Estimators which result from minimizing (2.1) are thus those which minimize the integrated squared moduli of residuals defined by the Fourier transforms in (2.2).

Explicit integration of (2.1) yields

$$ {}^{-\frac{1}{2}p}[n^{-2}(m^2)^{-\frac{1}{2}p}\sum_j\sum_k \exp\{-(4m^2)^{-1}(\underset{\sim}{y}_j - \underset{\sim}{y}_k)^T(\underset{\sim}{y}_j - \underset{\sim}{y}_k)\}$$

$$- 2\ n^{-1}(m^2 + \tfrac{1}{2})^{-\frac{1}{2}p}\sum_j \exp\{-(2 + 4m^2)^{-1}\underset{\sim}{y}_j^T\underset{\sim}{y}_j\}$$

$$+ (m^2 + 1)^{-\frac{1}{2}p}] \quad , \qquad (2.3)$$

where

$$y_j = D^{-\frac{1}{2}}(x_j - \mu) \quad .$$

Equating the partial derivatives of (2.3) with respect to μ and D (Dwyer, 1967) to zero gives the estimating equations for μ and D as, respectively

$$n^{-1}\pi^{\frac{1}{2}p}(m^2 + \tfrac{1}{2})^{-\frac{1}{2}p-1} \sum_j D^{-1}(x_j - \mu)\exp\{-(4m^2 + 2)^{-1}(x_j - \mu)^T D^{-1}(x_j - \mu)\} = 0 \ , \tag{2.4}$$

and

$$\pi^{\frac{1}{2}p}[-\tfrac{1}{2}n^{-2}(m^2)^{-\frac{1}{2}p-1} \sum_j \sum_k D^{-1}(x_j - x_k)(x_j - x_k)^T D^{-1}$$

$$\exp\{-(4\,m^2)^{-1}(x_j - x_k)^T\,D^{-1}(x_j - x_k)\}$$

$$+ \tfrac{1}{2}n^{-1}(m^2 + \tfrac{1}{2})^{-\frac{1}{2}p-1} \sum_j D^{-1}(x_j - \mu)(x_j - \mu)^T D^{-1}$$

$$\exp\{-(4\,m^2 + 2)^{-1}(x_j - \mu)^T\,D^{-1}(x_j - \mu)\}] = 0 \quad . \tag{2.5}$$

From (2.4) and (2.5) we find that the estimators for μ and D, $\hat{\mu}_1$ and $\hat{D}_1$, say, satisfy the implicit equations

$$\mu = \frac{\sum_j x_j v_{1jm}}{\sum_j v_{1jm}} \ , \tag{2.6}$$

and

$$A = B \ , \tag{2.7}$$

where

$$\underset{\sim}{A} = (m^2 + \tfrac{1}{2})^{-\frac{1}{2}p-1} \sum_j (\underset{\sim}{x}_j - \underset{\sim}{\mu})(\underset{\sim}{x}_j - \underset{\sim}{\mu})^T$$

$$\times \exp\{-(4m^2 + 2)^{-1}(\underset{\sim}{x}_j - \underset{\sim}{\mu})^T \underset{\sim}{D}^{-1}(\underset{\sim}{x}_j - \underset{\sim}{\mu})\} , \qquad (2.8)$$

and

$$\underset{\sim}{B} = (2n)^{-1}(m^2)^{-\frac{1}{2}p-1} \sum_j \sum_k (\underset{\sim}{x}_j - \underset{\sim}{x}_k)(\underset{\sim}{x}_j - \underset{\sim}{x}_k)^T$$

$$\times \exp\{-(4m^2)^{-1}(\underset{\sim}{x}_j - \underset{\sim}{x}_k)^T \underset{\sim}{D}^{-1}(\underset{\sim}{x}_j - \underset{\sim}{x}_k)\} , \qquad (2.9)$$

and

$$v_{1jm} = \exp\{-(4\ m^2 + 2)^{-1}(\underset{\sim}{x}_j - \underset{\sim}{\mu})^T\ \underset{\sim}{D}^{-1}(\underset{\sim}{x}_j - \underset{\sim}{\mu})\} . \qquad (2.10)$$

The generalized distances $(\underset{\sim}{x}_j - \underset{\sim}{\mu})^T \underset{\sim}{D}^{-1}(\underset{\sim}{x}_j - \underset{\sim}{\mu})$ and $(\underset{\sim}{x}_j - \underset{\sim}{x}_k)^T\ \underset{\sim}{D}^{-1}(\underset{\sim}{x}_j - \underset{\sim}{x}_k)$ constitute an essential component of the estimation process defined by (2.6)-(2.10). These distances have been used in a variety of contexts including clustering (Gnanadesikan, 1977, Ch. 4), identification of potential outliers (Barnett and Lewis, 1978, Ch. 6; Gnanadesikan and Kettenring, 1972; Rohlf, 1975), and regression diagnostics (Belsley, Kuh, and Welsch, 1980, Ch. 2). Because the diagnostic capability of these generalized distances is embodied in the estimation process itself, we have in a real sense combined the processes of estimation and criticism. The word criticism is used here in the sense of Daniel (1959), Box (1979, p. 2), and Paulson and Nicklin (1983). Criticism is the process of assessing the internal consistency of the data and the tentative model, in this case that the $\underset{\sim}{x}_j$ are independently distributed according to a p-variate Gaussian law.

As $m \to \infty$, the estimator $\hat{\mu}_1 \to n^{-1} \sum x_j$, the ordinary sample mean vector. The covariance matrix D is not strictly estimable through (2.7) when $m \to \infty$ since we end up with, at $\hat{\mu}_1 = \mu$, the identity

$$\frac{1}{n-1} \sum_j (x_j - \hat{\mu}_1)(x_j - \hat{\mu}_1)^T = \frac{1}{2\,n(n-1)} \sum_j \sum_k (x_j - x_k)(x_j - x_k)^T \tag{2.11}$$

which does not explicitly involve D. Either side of (2.11) may be used as an estimator for D on the truth of the independent, identical, p-variate normality of the x_j. The estimators $\hat{\mu}_1$ and $\hat{D}_1$ are not m-estimators because of their explicit dependence on the pairwise generalized distances. The comment that either side of (2.11) may be used to produce estimators for D leads us to presently develop two different estimators for D.

3. SIMPLIFIED ESTIMATORS OF $\underset{\sim}{D}$

The estimator $\hat{\underset{\sim}{D}}_1$ is numerically difficult to extract from (2.7), especially if p is large. Fixed point and Newton-Raphson algorithms for producing $\hat{\underset{\sim}{D}}_1$ have been developed and found to work well except for computational expense. However, the intriguing identity (2.7) at the solution $\hat{\underset{\sim}{\mu}}_1$ and $\hat{\underset{\sim}{D}}_1$ suggest two alternative estimators for $\underset{\sim}{D}$. The first is based on $\underset{\sim}{A}$ of (2.8). The joint estimators for $\underset{\sim}{\mu}$ and $\underset{\sim}{D}$ are obtained by determining the value K_2 in

$$\underset{\sim}{S}_2 = \sum_j \{K_2(\underset{\sim}{x}_j - \underset{\sim}{\mu})(\underset{\sim}{x}_j - \underset{\sim}{\mu})^T - \underset{\sim}{D}\}\exp\{-(4m^2 + 2)^{-1}(\underset{\sim}{x}_j - \underset{\sim}{\mu})^T\underset{\sim}{D}^{-1}(\underset{\sim}{x}_j - \underset{\sim}{\mu})\} \tag{3.1}$$

which makes $E(\underset{\sim}{S}_2) = \underset{\sim}{0}$. Direct computation gives

$$K_2 = \frac{2 + 2m^2}{1 + 2m^2} \quad . \tag{3.2}$$

Thus the joint estimators for $\underset{\sim}{\mu}$ and $\underset{\sim}{D}$, $\hat{\underset{\sim}{\mu}}_2$ and $\hat{\underset{\sim}{D}}_2$, satisfy the implicit equations

$$\underset{\sim}{\mu} = \frac{\sum_j \underset{\sim}{x}_j \, v_{2jm}}{\sum_j v_{2jm}} \, , \tag{3.3}$$

$$\underset{\sim}{D} = \frac{2 + 2m^2}{1 + 2m^2} \; \frac{\sum_j (\underset{\sim}{x}_j - \underset{\sim}{\mu})(\underset{\sim}{x}_j - \underset{\sim}{\mu})^T v_{2jm}}{\sum_j v_{2jm}} \, , \tag{3.4}$$

with

$$v_{2jm} = \exp\{-(4m^2 + 2)^{-1}(\underset{\sim}{x}_j - \underset{\sim}{\mu})^T\underset{\sim}{D}^{-1}(\underset{\sim}{x}_j - \underset{\sim}{\mu})\} \quad . \tag{3.5}$$

Another pair of estimators for $\underset{\sim}{\mu}$ and $\underset{\sim}{D}$ may be based on $\underset{\sim}{B}$ of the right hand side of equation (2.8). We require the constant K_3 in

$$\underset{\sim}{S}_3 = \sum_{\substack{j \\ j \neq k}} \sum_{k} \{K_3(\underset{\sim}{x}_j - \underset{\sim}{x}_k)(\underset{\sim}{x}_j - \underset{\sim}{x}_k)^T - \underset{\sim}{D}\} \exp\{-4m^2)^{-1}(\underset{\sim}{x}_j - \underset{\sim}{x}_k)^T \underset{\sim}{D}^{-1}(\underset{\sim}{x}_j - \underset{\sim}{x}_k)\} \tag{3.6}$$

which makes $E(\underset{\sim}{S}_3) = \underset{\sim}{0}$. Since $\underset{\sim}{x}_j - \underset{\sim}{x}_k$, $j \neq k$, is $N_p(\underset{\sim}{0}, 2\underset{\sim}{D})$ we find by direct computation that

$$K_3 = \frac{1}{2(n-1)} (1 + \frac{1}{m^2}) \{(n-1) + (1 + \frac{1}{m^2})^{\frac{1}{2}p}\} \quad . \tag{3.7}$$

Thus the joint estimators for $\underset{\sim}{\mu}$ and $\underset{\sim}{D}$, $\hat{\underset{\sim}{\mu}}_3$ and $\hat{\underset{\sim}{D}}_3$, satisfy the implicit equations

$$\underset{\sim}{\mu} = \frac{\sum\sum_{j \neq k} \underset{\sim}{x}_j \, v_{3jkm}}{\sum\sum_{j \neq k} v_{3jkm}} , \tag{3.8}$$

$$\underset{\sim}{D} = K_3 \frac{\sum\sum_{j \neq k} (\underset{\sim}{x}_j - \underset{\sim}{x}_k)(\underset{\sim}{x}_j - \underset{\sim}{x}_k)^T v_{3jkm}}{\sum\sum_{j \neq k} v_{3jkm}} , \tag{3.9}$$

with

$$v_{3jkm} = \exp\{-(4m^2)^{-1}(\underset{\sim}{x}_j - \underset{\sim}{x}_k)^T \underset{\sim}{D}^{-1}(\underset{\sim}{x}_j - \underset{\sim}{x}_k)\} \quad . \tag{3.10}$$

Both estimators $\hat{\underset{\sim}{D}}_2$ and $\hat{\underset{\sim}{D}}_3$ are computationally more attractive than $\hat{\underset{\sim}{D}}_1$. All of $\hat{\underset{\sim}{D}}_j$, $j = 1, 2, 3$, are positive definite with probability one for $m > 0$ whenever $n > p$ although all may be algorithmically singular. The estimator $\hat{\underset{\sim}{D}}_2$ is an M estimator for $\underset{\sim}{D}$ but $\hat{\underset{\sim}{D}}_1$ and $\hat{\underset{\sim}{D}}_3$ are not because of their explicit dependence on the pairwise generalized distances. All of $\hat{\underset{\sim}{D}}_j$, $j = 1, 2, 3$ are affine invariant. The forms of the equations (2.7), (3.1), and (3.6) imply that the influence of a single observation is bounded and re-descendent so that the estimators are qualitatively robust. The weights v_{1jm} , v_{2jm} , and v_{3jm} , $0 < m < \infty$, provide useful diagnostics for assessing the character of the data vis-a-vis the normality assumption. These weights are also useful in identifying potential outliers or data which require further study. Overall patterns in the final weights are also useful in this regard. These points are illustrated in a subsequent example.

The asymptotic efficiency of the estimators $\hat{\underset{\sim}{\mu}}_j$ and $\hat{\underset{\sim}{D}}_j$ relative to the maximum likelihood estimators is an increasing function of m and approaches 1 as $m \to +\infty$. The efficiencies of $\hat{\underset{\sim}{\mu}}_2$ and $\hat{\underset{\sim}{D}}_2$ among the three pairs of estimators are easiest to obtain.

The efficiency of the j^{th} component of $\hat{\underset{\sim}{\mu}}_2$, say $\hat{\mu}_{2j}$, relative to the sample mean is easily determined and is given by

$$\text{eff}(\hat{\mu}_{2j}) = \left(1 + \frac{c^2}{1 + 2c}\right)^{-\frac{1}{2}(p+2)} ; \tag{3.11}$$

the efficiency of the j^{th} diagonal component of $\hat{\underset{\sim}{D}}_2$, say d_{2j}^2 , relative to the usual maximum likelihood estimator is only a little more difficult to obtain and it is given by

$$\text{eff}(d_{2j}^2) = \left(1 + \frac{3c^2}{2 + 4c}\right)^{-1} \left(1 + \frac{c^2}{1 + 2c}\right)^{-\frac{1}{2}(p+2)} , \tag{3.12}$$

where $c = (1 + 2m^2)^{-1}$.

Figure 1 provides some of these efficiencies as a function of m and p . The efficiencies associated with $\hat{\underset{\sim}{\mu}}_3$ ($\hat{\underset{\sim}{\mu}}_1$ has the same asymptotic efficiency as $\hat{\underset{\sim}{\mu}}_2$) and $\hat{\underset{\sim}{D}}_3$ are approximately those of $\hat{\underset{\sim}{\mu}}_2$ and $\hat{\underset{\sim}{D}}_2$. A small Monte Carlo study suggests that the efficiencies related to $\hat{\underset{\sim}{D}}_1$ are slightly higher than those of $\hat{\underset{\sim}{D}}_2$.

The efficiencies of the off-diagonal element of $\hat{\underset{\sim}{D}}_2$ relative to the maximum likelihood estimates of covariances are very nearly equal to those of the diagonal elements.

The contours of all influence functions of the estimators $\hat{\underset{\sim}{\mu}}_j$, $\hat{\underset{\sim}{D}}_j$ at the multivariate normal are all closed and bounded. This closedness property implies that estimation may be combined with clustering or may be used as a clustering algorithm or used to evaluate the results of a clustering algorithm. The clustering capability associated with estimation of $\underset{\sim}{\mu}$ and $\underset{\sim}{D}$ allows for an identification of potential outliers in multivariate normal data.

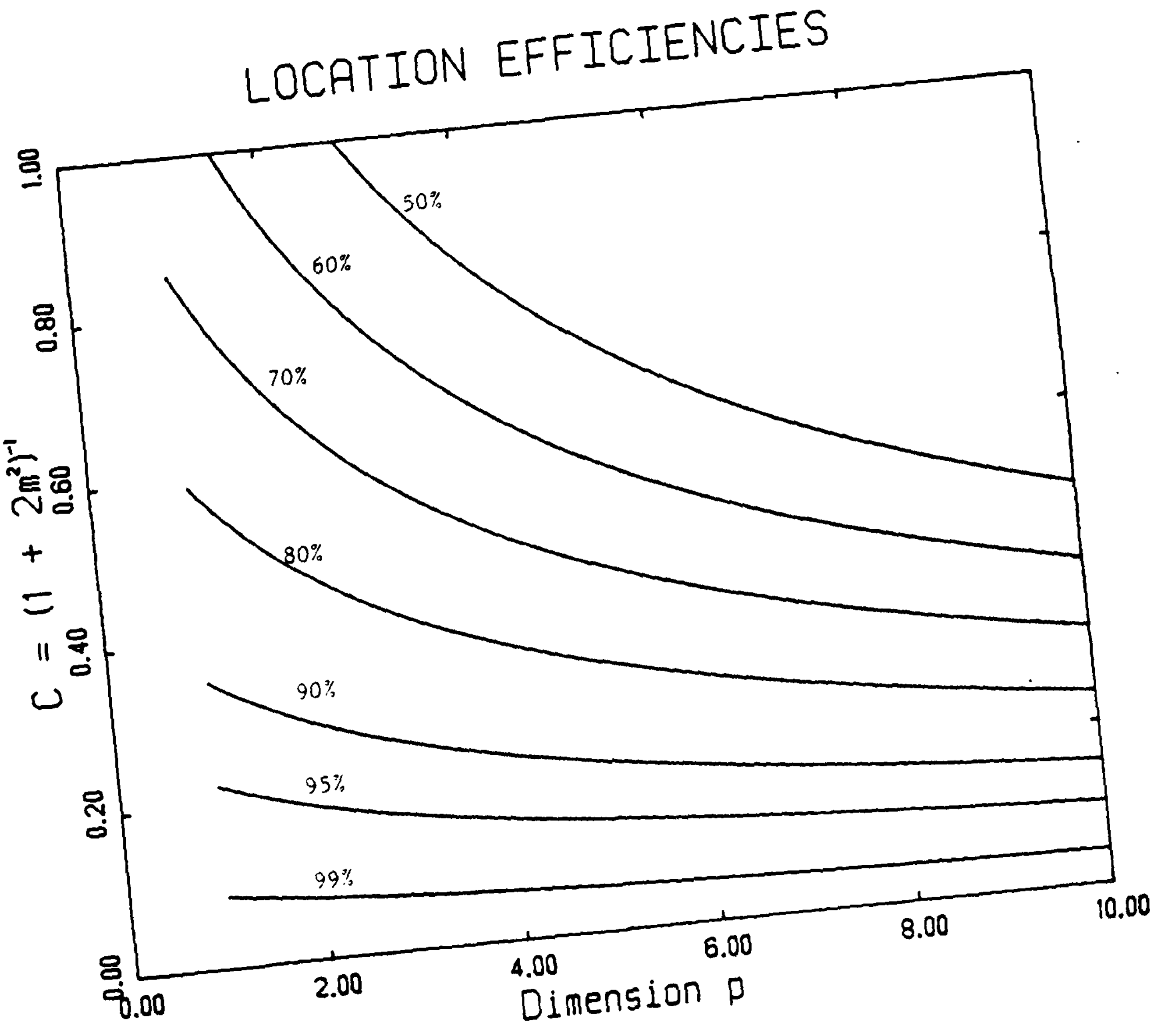


FIGURE 1a. Efficiencies of $\hat{\mu}_{2j}$ as a function of m and p .

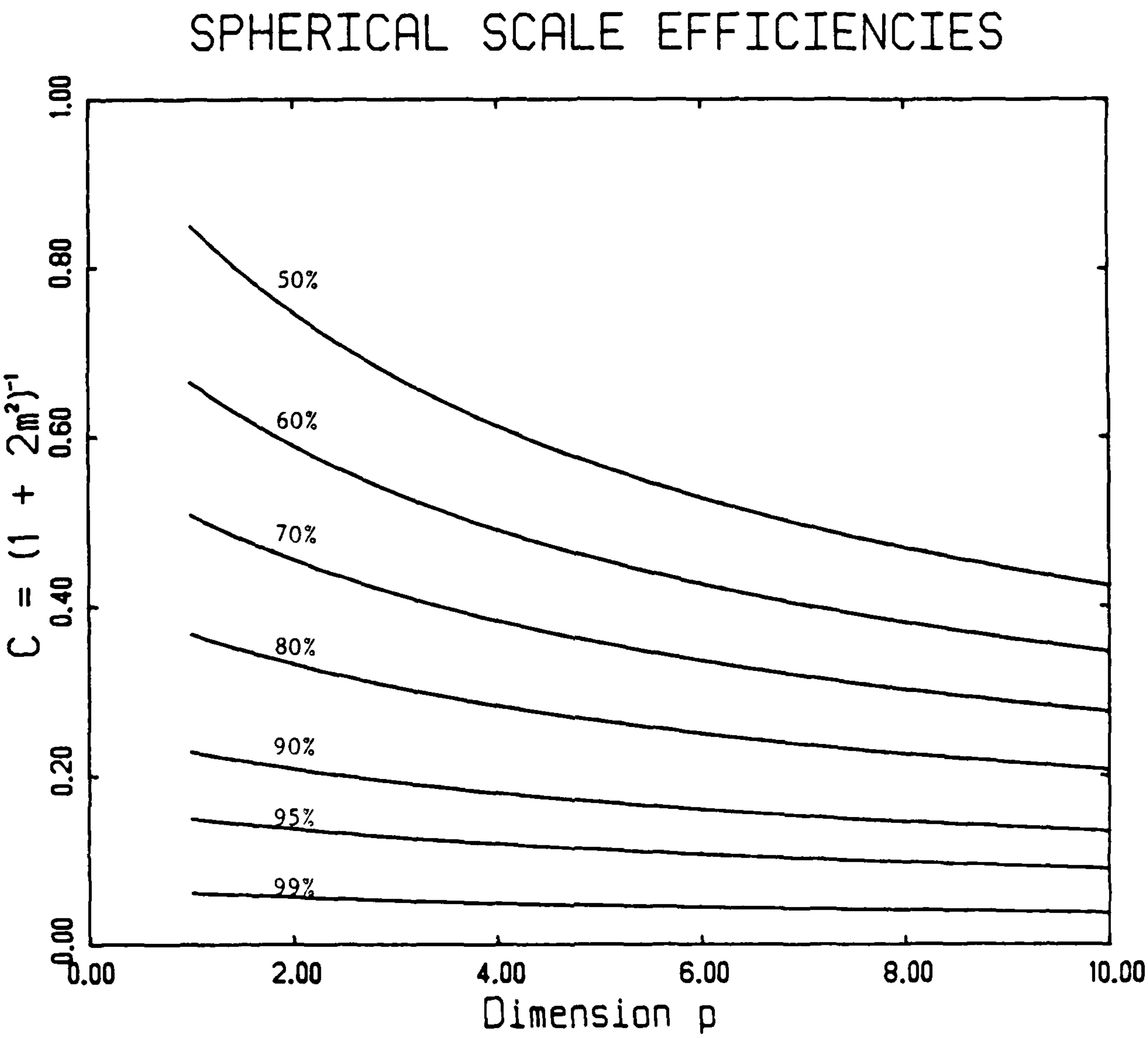


FIGURE 1b. Efficiencies of $\hat{d}^2_{2j}$ as a function of m and p.

4. PARAMETRIC DENSITY ESTIMATION

The estimators $\hat{\mu}_1$ and $\hat{D}_1$ of Section 2 are equivalent to those derived from the following parametric density estimation considerations. The expression $R(u)$ of (2.2) represents a difference of characteristic functions whose inverse is

$$r(x) = g(x) - g_n(x) \tag{4.1}$$

where $g(x)$ is the spherical normal distribution with

$$g(x) = |2\pi(1 + m^2)I|^{-\frac{1}{2}}\exp\{-\tfrac{1}{2}(1 + m^2)^{-1}x^T x\} \tag{4.2}$$

and $g_n(x)$ is an estimator of $g(x)$ with

$$g_n(x) = n^{-1} \sum_j |2\pi m^2 I|^{-\frac{1}{2}}\exp\{-(2m^2)^{-1} z_j^T z_j\} \;, \tag{4.3}$$

$$z_j = (x - D^{-\frac{1}{2}}(x_j - \mu)) \quad . \tag{4.4}$$

The matrix I is the $p\times p$ identity matrix. The expression $g_n(x)$ is unbiased for $g(x)$ when the x_j are p-variate Gaussian. By the multidimensional version of Parseval's theorem

$$Q_{m,n}(\mu, D) = \int_{R_p} |R(u)|^2 du = (2\pi)^p \int_{R_p} R^2(x)dx \tag{4.5}$$

Accordingly, the estimators $\hat{\mu}_1$ and $\hat{D}_1$ are those arrived at from minimizing the integrated squared residual in x and may be accurately termed parametric density estimators.

The robustness characteristics of the estimators $\hat{\mu}_1$ and $\hat{D}_1$ are easily seen from working with (4.5), e.g., differentiation of (4.5) with respect to $\underset{\sim}{\mu}$ gives

$$\int_{R_p} \frac{\partial\, g_n(\underset{\sim}{x})}{\partial\, \underset{\sim}{\mu}} (g(\underset{\sim}{x}) - g_n(\underset{\sim}{x}))d\underset{\sim}{x} = 0 \quad .$$

From this equation we readily deduce for $0 < m < \infty$ that the marginal contribution of $\underset{\sim}{x}_n$, say, to the estimate of $\underset{\sim}{\mu}$ depends on all the $\underset{\sim}{x}_j$, not just $\underset{\sim}{x}_n$, as well as the assumed p-variate Gaussian density, that it is bounded, and that it is redescendent as $||x_n|| \to \infty$. The function $g_n(\underset{\sim}{x})$ is similar to a Parzen kernel density estimator; the form of the kernel is determined by the function $\exp(-m^2\underset{\sim}{u}^T\underset{\sim}{u})$ of (2.1). The advantage to the choice of the Gaussian kernel is that it leads to convenient computational expressions. Choice of functions $|\omega(\underset{\sim}{u})|^2$ other than $\exp(-m^2\underset{\sim}{u}^T\underset{\sim}{u})$ generally leads to kernels which do not permit closed form integration of (4.5) and thus make for intractable numerics when $p > 2$. Intractable numerics result from the need to compute integrals by numerical methods.

Indeed, the potential for application of parametric density estimation procedures is bleak if integrals of the form (2.1) or (4.5) for a non-Gaussian p-variate density $f(\underset{\sim}{x})$ cannot be evaluated in closed form. The reason is simple: high order numerical integration for p large is not viable even for today's computers.

5. AN EXAMPLE AND DISCUSSION

A choice of m is needed to implement the estimators of Sections 2 and 3. We feel that exploration of the data vis-a-vis the p-variate normal model is in many cases more useful since it leads to a deeper understanding of the data generating process and the actual problem at hand. Accordingly, it will often be useful to use a range of values of m in our analyses since the responses of the parameter estimates and final weights to changes in m generates very useful diagnostics.

Table 1 gives the chemical analysis of 20 geological specimens in terms of percentage of iron (Fe), sodium (Na), and potassium (K) which had been believed by geologists to be approximately homogeneous in character and also trivariate Gaussian. All two way scatterplots of this data are given in Figure 2. A few points visually "stick out" enough to question the belief of homogeneity and simple 3-variate Gaussianity. It is of interest, however, to determine how the three robust estimators of covariance react to this data for various values of m.

Table 2 presents the estimates of the variances and correlations for $m^2 = 10, 5, 2$ and maximum likelihood. At $m^2 = 10$, all estimates are reasonably close to those of maximum likelihood and at $m^2 = 5$ there are some minor differences. However, at $m^2 = 2$ we find that there are now important differences in the estimates. These differences provide a warning that something may be awry regarding the internal consistency of the data and the assumption of 3-variate normality. Of particular interest are the estimates of the correlation ρ_{23}

TABLE 1

CHEMISTRY OF GEOLOGICAL SPECIMENS

Observation Number	Fe	Na	K
A	2.6	1.7	3.4
B	2.1	2.1	3.4
C	1.3	2.8	1.7
D	2.2	2.1	3.0
E	1.3	2.2	3.6
F	1.6	2.2	3.7
G	3.1	2.2	2.4
H	2.8	1.7	3.8
I	4.0	1.4	3.3
J	2.6	1.9	3.2
K	1.5	2.0	4.2
L	3.9	1.6	2.5
M	3.1	1.7	3.5
N	3.1	1.9	3.6
O	1.7	2.1	3.6
P	1.4	2.2	4.0
Q	3.0	2.0	3.8
R	2.9	2.2	5.5
S	2.9	2.8	5.9
T	2.9	2.9	6.5

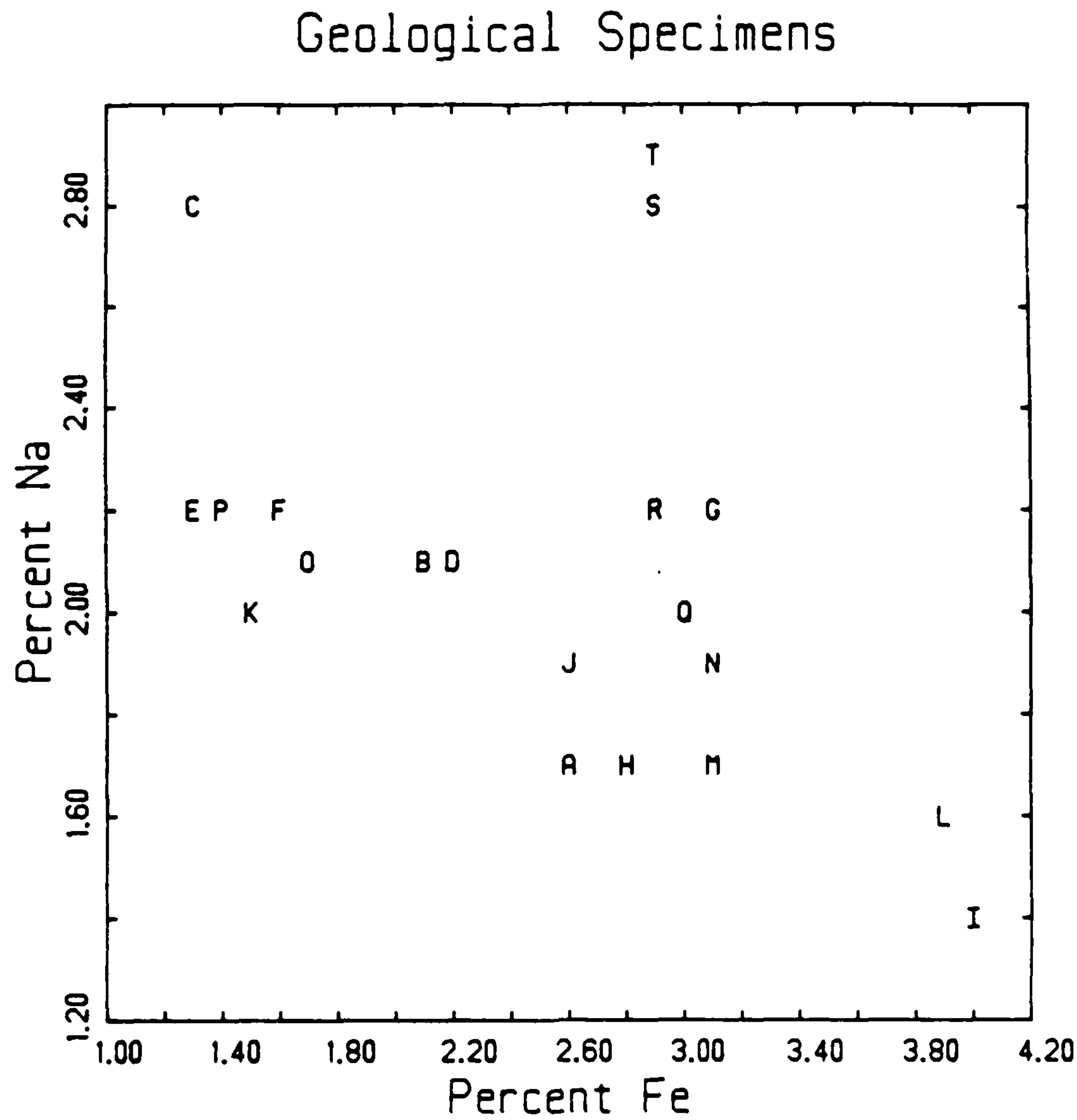


FIGURE 2a. Percentages of iron and sodium for geological specimens.

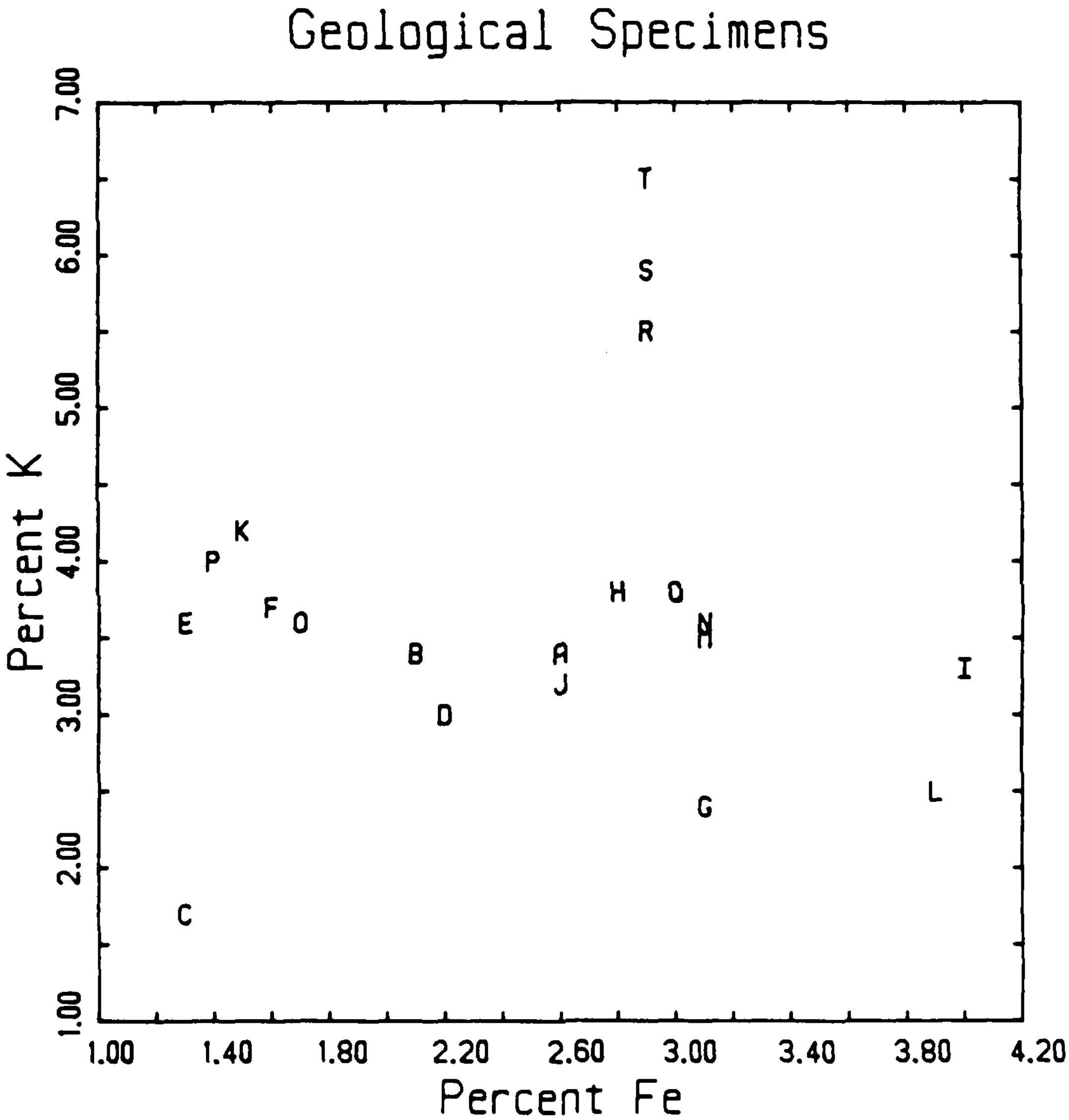


FIGURE 2b. Percentages of iron and potassium for geological specimens.

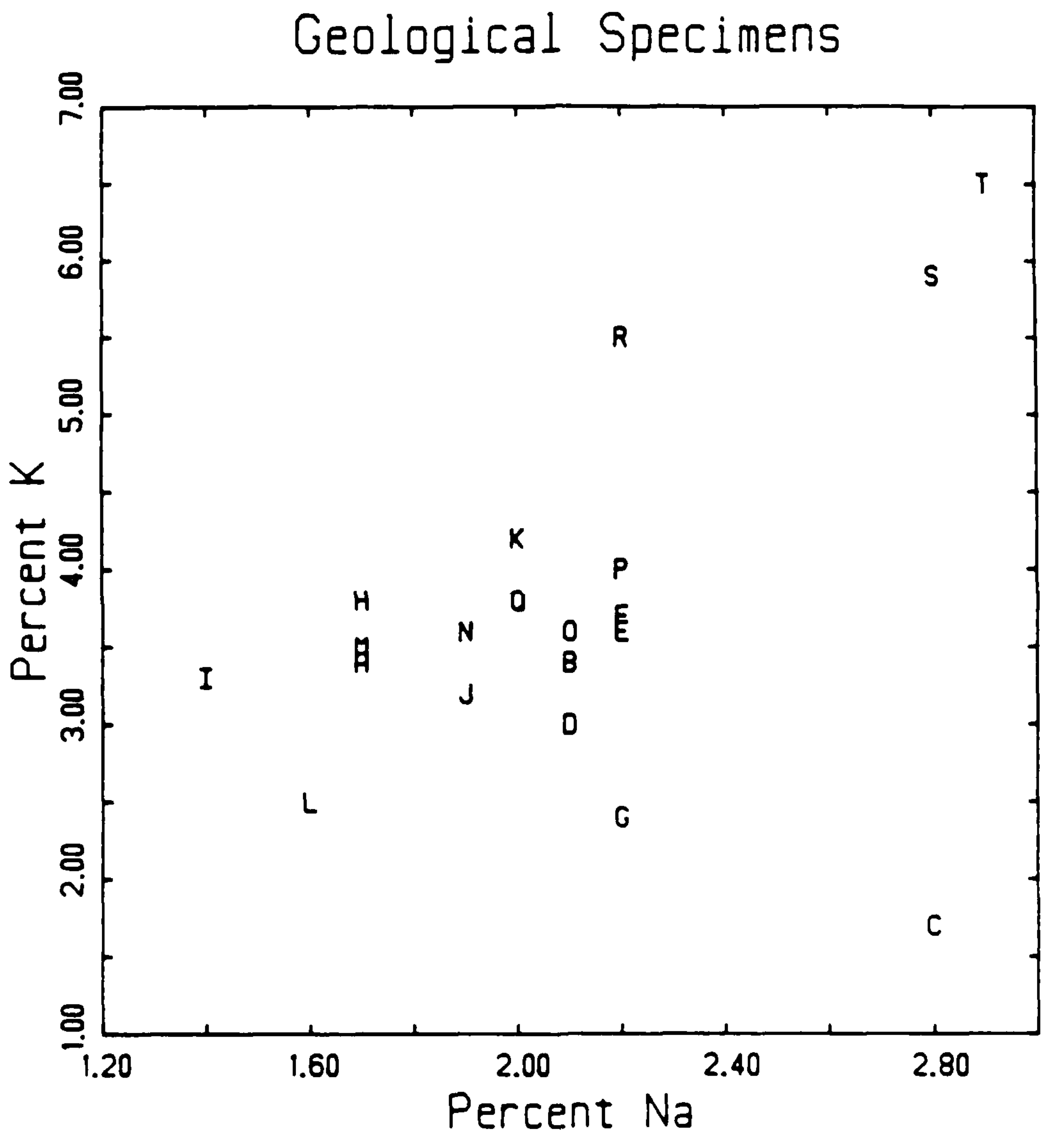


FIGURE 2c. Percentages of sodium and potassium for geological specimens.

TABLE 2

PARAMETER ESTIMATES FOR SEVERAL VALUES OF m^2.

Estimator	m^2	$\hat{d}_i^2$ Fe	Na	K	$\hat{\rho}_{12}$	$\hat{\rho}_{13}$	$\hat{\rho}_{23}$
D_1	10	0.653	0.151	1.26	-0.419	0.035	0.441
	5	0.651	0.152	1.30	-0.396	0.085	0.473
	2	1.19	0.166	0.633	-0.770	-0.133	-0.413
D_2	10	0.659	0.143	1.16	-0.438	0.058	0.438
	5	0.671	0.135	1.07	-0.446	0.021	0.464
	2	0.776	0.070	0.229	-0.824	-0.508	0.137
D_3	10	0.701	0.153	1.23	-0.439	0.059	0.436
	5	0.722	0.144	1.14	-0.450	0.020	0.458
	2	0.828	0.075	0.245	-0.828	-0.496	0.139
MLE		0.646	0.149	1.23	-0.435	0.089	0.416

for the four estimators. Maximum likelihood perceives a relatively strong positive correlation between Na and K, the estimators $\hat{D}_2$ and $\hat{D}_3$ perceives a relatively weak positive correlation between Na and K, and the estimator $\hat{D}_1$ perceives a relatively strong negative correlation between Na and K. An examination of Figure 1c leads one to conclude that each interpretation is permissible, depending on which data points are considered inconsistent with the remainder of the data and the 3-variate Gaussianity. It should be emphasized that a purely robust procedure which focused on 95% efficiency may not have picked up a potential problem with the data-model unit while an exploratory (sensitivity of solutions to changes in m) analysis did. Of course, for this 3-dimensional situation we can easily see by graphical methods that the data are not homogeneous and that there may be some interesting geological structure which warrants further investigation. Since simple graphical methods may prove less effective in higher dimensions, there is a real advantage to using these estimators in combination with graphical methods for exploration and deeper understanding of data and its generation.

The final weights $\hat{v}_{\ell jm}$ and $\hat{w}_{\ell jkm}$, i.e., $v_{\ell jm}$ evaluated at $\hat{\mu}_\ell$, $\hat{D}_\ell$, are useful in diagnostic examination of the data. Table 3 presents the final observational weights for $\tilde{m} = 2$, $\hat{v}_{\ell jm}$ for $\ell = 1, 2$ and $\sum_{k \neq j} \hat{v}_{3jkm}$ for $\hat{D}_3$. All three estimation procedures call attention to observations 18, 19, and 20 as being potentially different from the remainder in the assumption of Guassianity. Methods 2 and 3 strongly indicate that observation 3 is potentially different while the estimator 1 weakly indicates the potential difference. We do

TABLE 3

FINAL OBSERVATIONAL WEIGHTS FOR $m^2 = 2$.

Observation Number	$\hat{D}_1$ / $m^2 = 2$	$\hat{D}_2$ / $m^2 = 2$	$\hat{D}_3$ / $m^2 = 2$
1	0.058	0.061	0.059
2	0.069	0.078	0.074
3	0.038	0.005	0.011
4	0.068	0.069	0.066
5	0.062	0.067	0.064
6	0.065	0.073	0.070
7	0.051	0.033	0.034
8	0.067	0.069	0.066
9	0.054	0.052	0.051
10	0.068	0.077	0.074
11	0.060	0.062	0.059
12	0.049	0.048	0.047
13	0.067	0.073	0.070
14	0.056	0.055	0.055
15	0.067	0.075	0.072
16	0.060	0.067	0.064
17	0.042	0.036	0.038
18	0.6×10^{-3}	0.5×10^{-4}	0.007
19	0.1×10^{-6}	0.3×10^{-9}	0.010
20	1.1×10^{-8}	0.3×10^{-12}	0.010

not know which, if any, of the data points are really different, but we do want to know of the potential existence of such points since this information is useful in evaluating our models and our knowledge of the phenomenon under examination.

REFERENCES

Barnett, V., and Lewis, T. (1978). Outliers in Statistical Data. New York: Wiley.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York: Wiley.

Box, G. E. P. (1979). Some problems of statistics and everyday life. J. Amer. Statist. Ass., 74, 1-4.

Daniel, C. (1959). Use of half-normal plots in interpreting factorial two-level experiments. Technometrics, 1, 311-341.

Dwyer, P. S. (1967). Some applications of matrix derivatives in multivariate analysis. J. Amer. Statist. Ass., 62, 607-625.

Gnanadesikan, R. (1977). Methods for Statistical Data Analysis of Multivariate Observations. New York: Wiley.

Gnanadesikan, R., and Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. Biometrics, 28, 81-124.

Parzen, E. (1962). On estimation of a probability density and mode. Ann. Math. Statist., 33, 1065-1076.

Paulson, A. S., and Nicklin, E. H. (1983). Integrated distance estimators for linear models, applied to some published data sets. To appear in Applied Statistics.

Rohlf, F. J. (1975). Generalization of the gap test for the detection of multivariate outliers. Biometrics, 31, 93-101.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. Ann. Math. Statist., 27, 832-837.

Watson, G. S., and Leadbetter, M. R. (1963). On the estimation of the probability density, I. Ann. Math. Statist., 34, 480-491.

MASTER COPY - FOR REPRODUCTION PURPOSES

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
ARO-4	N/A	N/A

4. TITLE (and Subtitle)
Some Robust Estimates of Covariance Structure Based on Parametric Density Estimation

5. TYPE OF REPORT & PERIOD COVERED
Working Paper

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
A.S. Paulson
T.A. Delehanty
N.J. Delaney

8. CONTRACT OR GRANT NUMBER(s)
DAAG29-81-K-0110

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Rensselaer Polytechnic Institute
Troy, New York 12180

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS
U. S. Army Research Office
Post Office Box 12211
Research Triangle Park, NC 27709

12. REPORT DATE

13. NUMBER OF PAGES

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)
Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)
NA

18. SUPPLEMENTARY NOTES
The view, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
robust estimators, covariance matrices, generalized likelihood, re-descendent functions, critical estimation procedures, parametric density estimation

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)
A new family of Fourier-based estimators of the parameters of the multivariate Gaussian distribution is presented. The estimators are equivalent to parametric density estimators. Three distinct estimators arise, each of which is robust and reduces to the maximum likelihood estimator as a special case. By varying the window width of a parametric density estimator, a set of diagnostics which are useful in problems of outlier detection and clustering are obtained. An example, using a trivariate data set, is given.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

END

4-87

DTIC